Metadata Discovery and Integration to Support Repurposing of Heterogeneous Data using the OpenFurther Platform

bioCADDIE All Hands Meeting
September 11th, 2016

Ram Gouripeddi
&
Julio Facelli
Overview

• Background
• Goals
• Tasks & Updates
• Next Steps
OpenFurther - Complex Integration
Current State of Art

- Relies on human manual curation
- Non-scalable
- Limits realization of full potential of using big data technologies in biomedical science.

<table>
<thead>
<tr>
<th>Data Stream</th>
<th>Metadata</th>
<th>Terminology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laboratory</td>
<td>228</td>
<td>6848</td>
</tr>
<tr>
<td>Microbiology</td>
<td>504</td>
<td>2793</td>
</tr>
<tr>
<td>Radiology</td>
<td>156</td>
<td>2381</td>
</tr>
<tr>
<td>Total</td>
<td>888</td>
<td>12022</td>
</tr>
</tbody>
</table>

Final number of metadata and semantic elements discovered, mapped and stored in OpenFurther for initial onboarding of laboratory, microbiology and radiology results for 6 hospitals in the PHIS+ network. (Numbers do not include artifacts that were considered as potential matches in the preceding discovery phase.)
OpenFurther - Complex Integration

Data Services
- Discovery, Search & Retrieval
- Federation & Aggregation
- Curation
- Data Publication
- Linkage with Literature

Metadata & Semantics – Discovery & Mapping Service

Automating/Semi-automating

www.openfurther.org
Goals

Prototype a computational infrastructure that supports discovery and mapping of metadata artifacts and terminologies to overcome current limitations.

Automated/semi-automated depending on confidence threshold chosen at specific implementations.

Agnostic to specific algorithms or tools as many of these are domain-specific and dependent on data.

Choose best available solution(s) based on performance, and for specific metadata artifacts.

Leverage existing components of OpenFurther framework to manage, integrate and share metadata in structured non-proprietary formats.
Tasks

1. Catalogue, benchmark and evaluate existing metadata discovery and mapping methods and tools.

2. Adapting bioCADDIE metadata specifications into OpenFurther’s Metadata Repository.


Catalogue, Benchmark and Evaluate existing Metadata Discovery and Mapping Methods and Tools
# Metadata Mapping/Matching Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarity Flooding</td>
<td>- Uses similarity propagation</td>
</tr>
<tr>
<td>Artemis (Analysis of Requirements: Tool Environment for Multiple Information)</td>
<td>- Affinity-based analysis and hierarchical clustering of source schema elements</td>
</tr>
<tr>
<td>Cupid</td>
<td>- Computes similarity coefficients with a domain specific dictionary</td>
</tr>
<tr>
<td>COMA (COMbination of MAtrching algorithms)</td>
<td>- Multiple matching algorithms that can be evaluated against one another</td>
</tr>
<tr>
<td>Protoplasm</td>
<td>- Industrial strength matcher</td>
</tr>
<tr>
<td>S-Match</td>
<td>- Compares structures (e.g., XML schemas or ontologies) and returns semantic relations</td>
</tr>
<tr>
<td>Machine learning approaches</td>
<td></td>
</tr>
<tr>
<td>Combing indirect matches and direct matches</td>
<td></td>
</tr>
<tr>
<td>Multifaceted approaches</td>
<td></td>
</tr>
<tr>
<td>Distributed semantics and semantic indexing approaches</td>
<td></td>
</tr>
</tbody>
</table>

- Semantic matching methods and tools for analysis:
  - Metamap
  - MTERMS
  - cTAKES
  - LOINC mappers
Adapting bioCADDIE Metadata Specifications into OpenFurther’s Metadata Repository
OpenFurther’s Metadata Repository (MDR)

- Central to MDMS
- Built in-house, standards-based repository of artifacts and knowledge
- Artifacts include, but are not limited to
  - Logical models, local models, model mappings
  - Administrative information
  - Descriptive information
  - Translation Programs
- Conceptually, highly generic and abstracted entity relationship model
  - Assets: Things or elements within each artifact
  - Assets may have Properties and Associations to other assets.
  - Associations can have properties of their own and each association has a Type, which is also an asset.
- Can be implemented in relational or graph formats
- Stored metadata can be shared in various structured and non-proprietary formats (e.g. XML) using the translation programs
Conceptual Model of MDR

Schematic overview of the DATS core and extended elements, their types and relations.
DATS v2.1 (https://docs.google.com/document/d/1hYcYRlEkE6-dFm7qF9Bv1Oh51kTF6a8OhWUtvoZIDto/edit?usp=sharing), Alejandra Gonzalez-Beltran et al., (2016). WG3-MetadataSpecifications: DataMed DATS specification v2 - NIH DB2K bioCADDIE. Zenodo. 10.5281/zenodo.54010
Potential Benefits

• Flexible data resource metadata storage system
• Enhance the within resources metadata descriptors with descriptors of the resource itself.
• Versioning of model specifications
• Translation across different metadata specifications.
• Available for various data services.

Discovery, Search & Retrieval
Federation & Aggregation
Curation
Data Publication
Linkage with Literature
Conceptual Architecture of a metadata Discovery & Mapping Service (MDMS)
Key Considerations

• Framework to programmatically characterize metadata of a source.

• An extensible library of metadata discovery and mapping methods and tools.

• Standalone service oriented architecture.

• Use OpenFurther components (or equivalents) such as MDR, Ontology Service, Data Quality Framework

• MDMS accelerates multiple data services.
  • E.g. Data federation/integration with OpenFurther
High-level Workflow for Metadata Discovery

Determine File Type/Extensions

Learn Content Characteristics
- Structural pattern recognition
- Match with auxiliary descriptive (e.g. data dictionary)
- Store

Discover File using Content & Structure (Task 4)

Content of this kind of file Characterized?

Yes

No

Match using existing stores

New Dataset

Uncertainty

Human Interaction

Metadata Repository

Ontology Server

PDB

ClinVar

EPA

Clinical
Architectural Overview of MDMS
Machine Learning Methods for Metadata Discovery
Discover File using Structural & Content Metadata

- Text documents
  - Plain text
  - XML
  - HTML
  - Other

Tokenize

Document processing module

- Word Frequency Feature
- Word Structure Feature
- Other features

Document type knowledgebase (MDR)

Document Type Classifier

- Other features
Example Structural & Content Features

• Token frequency features
  • Token frequency
  • Normalized token frequency (normalized by length of document)

• Token structure features
  • Length of the word (median and mean)
  • Proportion of Capitalized Word
  • Proportion of words with ALL UPER LETTERS
  • Proportion of numerical value
  • Proportion of negative value

• Other features
  • Punctuation type and frequency : > _ : ?
  • Spacing
Preliminary Results Summary

Data

<table>
<thead>
<tr>
<th>Data Source</th>
<th>File Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCSB PDB</td>
<td>PDB</td>
</tr>
<tr>
<td>RCSB PDB</td>
<td>CIF</td>
</tr>
<tr>
<td>RCSB PDB</td>
<td>XML</td>
</tr>
<tr>
<td>Private protein structure (generated by I-TASSER)</td>
<td>PDB</td>
</tr>
<tr>
<td>ClinVar</td>
<td>Clinivar summary</td>
</tr>
<tr>
<td>ClinVar</td>
<td>XML</td>
</tr>
<tr>
<td>American National Corpus</td>
<td>Email</td>
</tr>
<tr>
<td>American National Corpus</td>
<td>Letter</td>
</tr>
</tbody>
</table>

Feature Selection

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Proportion of pure word</td>
</tr>
<tr>
<td>2</td>
<td>Proportion of UPPPER WORD</td>
</tr>
<tr>
<td>3</td>
<td>Proportion of numerical tokens</td>
</tr>
<tr>
<td>4</td>
<td>Proportion of other tokens (symbols and punctuations)</td>
</tr>
<tr>
<td>5</td>
<td>Median length</td>
</tr>
<tr>
<td>6</td>
<td>Normalized count of distinct word</td>
</tr>
<tr>
<td>7</td>
<td>Proportion of Capitalized Word</td>
</tr>
<tr>
<td>8</td>
<td>Proportion of negative numbers</td>
</tr>
</tbody>
</table>

- Accuracy: >90%
- AUC: >0.96
Next Steps

• Machine Learning work
  • Extend to environmental and clinical data
  • Conformity/Uncertainty

• Publish results on 4 tasks
  • Research Reproducibility 2016 Conference abstract accepted

• Share results with bioCADDIE and Core Development Team
  • September 27, 2016

• Continuation of works
  • RFA (e.g. NCATS Translator)
  • Collaborations & Opportunities with bioCADDIE and NIH BD2K
Thanks